

Eine fünfstufige Pipeline zur automatisierten Ausgewogenheitsanalyse öffentlich-rechtlicher Sendungen: Ein 150-Länder-Korpus und ein einheitlicher Analyserahmen

Methodik, Fassung 3.8

David Schläpfer

Schweizerischer Verein für ausgewogene Berichterstattung (SVFAB), Zürich, Schweiz

ORCID: 0009-0000-5671-9266 · Korrespondenzautor: kontakt@svfab.ch

Oktober 2026

Schlüsselwörter: Medienbias, öffentlich-rechtlicher Rundfunk, KI-gestützte Inhaltsanalyse, Ausgewogenheitsmessung, Quellenglaubwürdigkeit, Ländervergleich

Lizenz: CC BY-ND 4.0 (Namensnennung – Keine Bearbeitungen 4.0 International).

Kurzfassung in einfacher Sprache

Öffentlich-rechtliche Sender müssen per Gesetz sachgerecht und vielfältig berichten. Dieses Papier beschreibt, wie ein Computerprogramm Sendungen anhand ihrer Untertitel darauf prüft: Es zählt nach festen Regeln, wer zu Wort kommt, welche Begriffe fallen und was fehlt, und belegt jede Zählung mit Zitat und Zeitangabe. Abgeglichen wurde das Verfahren mit 159 Sendungen, über welche die Schweizer Beschwerdeinstanz UBI entschieden hat: Sendungen, bei denen sie eine Verletzung feststellte, erhalten im selben Sendeformat im Mittel höhere Werte. Das Verfahren folgt damit der Sicht der UBI: Zählt man die neun Fälle, in denen das Bundesgericht ihren Entscheid aufhob, zu den abgewiesenen, verschwindet der Unterschied weitgehend. Ein einzelner Wert schwankt zwischen zwei Durchläufen; verlässlich sind Aussagen über viele Sendungen.

Zusammenfassung

Dieses Papier stellt die vollständige Methodik eines Systems zur automatisierten Ausgewogenheitsanalyse öffentlich-rechtlicher Sendungsinhalte vor, angewandt auf ein archiviertes Korpus, das 150 Länder in 70 Sprachen abdeckt. Das System verarbeitet Untertitel (SRT/VTX, teils von den Sendern, teils aus eigener maschineller Transkription) von Nachrichten- und Informationssendungen mittels eines strukturierten Analyserahmens, der 5 Analyseschritte umfasst, 18 Kriterien — 17 Qualitätskriterien (K1–K18; K9 ist seit Fassung 4.0 aufgehoben, 16 tragen einen Score), die jeweils aus einer expliziten Zählregel und nicht aus einem holistischen Urteil abgeleitet werden und als eindeutig zählbar oder auslegungsbedürftig ausgewiesen sind, und als achtzehntes die politische Einordnung der Parteien (K19) —, die 6-dimensionale Quellenglaubwürdigkeitsbewertung (Quellenampel) als Werkzeug der Expertenauswahl, eine kontextabhängige Ausgewogenheitsbewertung sowie eine länderspezifische rechtliche Verankerung. Jede Sendung erhält einen Gesamtscore — das Mittel aller anwendbaren Kriterien — auf einer Skala von 0–10 (0 = neutral/ausgewogen, 10 = stark einseitig), ausgewiesen zusammen mit dem schwerwiegendsten Einzelbefund (Spitzenwert) und dem Mittelwert der drei schwerwiegendsten Befunde unter den Kriterien, die eine Regulierungsbehörde prüft (Schwere). Seit Fassung 3.6 weisen vier Kriterien zwei Werte aus: die von der schweizerischen Aufsichtsbehörde angewandte Messart und, an erster Stelle genannt, die Messart, die dieses Projekt für notwendig hält — Letztere ausdrücklich eine Position, kein Befund der Rechtswidrigkeit. Das Rahmenwerk ist an den Entscheiden der Unabhängigen Beschwerdeinstanz für Radio und Fernsehen (UBI) kalibriert und mit ihnen verglichen: In den 14 Sendeformaten, in denen beide

vorkommen, liegen die 34 Sendungen mit gutgeheissener Beschwerde im Mittel um 0.31 Punkte im Gesamtscore und um 0.66 Punkte in der Schwere über den 81 mit abgewiesener (Skala 0–10); in der Lesart des Bundesgerichts, das neun der Gutheissungen aufhob, verschwindet der Unterschied weitgehend. Die einzelne Sendung ordnet das Verfahren nicht zuverlässig zu. Die Regeln wurden an diesen Entscheiden entwickelt, eine Prüfung an Entscheiden, die bei der Entwicklung nicht vorlagen, steht aus. Zwei Durchläufe derselben Sendung weichen im Mittel um 0.71 Punkte im Gesamtscore und um 1.29 Punkte in der Schwere voneinander ab. Ein einziges universelles Prompt-Template, parametrisiert nach länderspezifischen Rechtsrahmen und politischen Landschaften, gewährleistet methodische Konsistenz über alle überwachten Länder hinweg. Dieser Bericht dient als kanonische methodische Referenz für alle Publikationen, Analysen und API-basierten Forschungszugänge, die darauf aufbauen.

1. Einleitung

Öffentlich-rechtliche Sender in demokratischen Gesellschaften unterliegen gesetzlichen Aufträgen, die eine ausgewogene, korrekte und pluralistische Berichterstattung verlangen. In der Schweiz kodifiziert Art. 4 RTVG diese Verpflichtung. Vergleichbare Bestimmungen bestehen in praktisch allen Ländern mit öffentlich-rechtlichen Rundfunksystemen: die BBC Charter (Vereinigtes Königreich), der Broadcasting Act (Kanada), Medieansvarslov (Norwegen), Zákon o ČT (Tschechische Republik), 放送法第4条 (Japan) und viele weitere.

Trotz dieser gesetzlichen Anforderungen fehlt eine systematische, kontinuierliche und vergleichbare Messung der Konformität einzelner Sendungen weitgehend. Bestehende Medienbeobachtung — sei es durch Regulierungsbehörden, akademische Institute oder Selbstregulierungsorgane — stützt sich typischerweise auf kleine Stichproben, manuelle Codierung durch kleine Teams und qualitative Bewertungen, die sich einer Quantifizierung und einem Ländervergleich entziehen. Das Ergebnis ist eine regulatorische Lücke: Gesetzliche Verpflichtungen bestehen, aber eine operative Messinfrastruktur fehlt.

Der hier beschriebene Rahmen schliesst diese Lücke durch eine vollständig automatisierte, KI-gestützte Pipeline, die Sendungstranskripte anhand von 18 Kriterien analysiert, strukturierte Berichte mit zeitgestempelten Belegen erstellt und vergleichbare Scores über Länder hinweg generiert. Neue Beiträge kommen jede Nacht hinzu.

Dieses Papier liefert die vollständige methodische Beschreibung des Systems auf Prompt-Fassung 4.14-detail (seit 27. September 2026). Es ersetzt frühere Beschreibungen in Begleitpapieren (Schlaepfer 2026a, 2026b, 2026c) und dient als kanonische Zitierquelle für Forschende, Regulierungsbehörden und Institutionen, die mit den vom System erzeugten Daten arbeiten.

2. Systemarchitektur

2.1 Überblick über die Pipeline

Die Pipeline verarbeitet Sendungsinhalte in fünf Stufen:

- Akquisition: Untertiteldateien (SRT/VTT) werden von öffentlich-rechtlichen Sendeplattformen gesammelt oder durch eigene maschinelle Transkription erzeugt und in zeitgestempelten Klartext mit [MM:SS]-Markierungen umgewandelt.
- Analyse: Jedes Transkript wird durch den universellen Analyse-Prompt (Abschnitt 3) verarbeitet, wobei ein strukturierter Bericht mit Scores, Belegen und rechtlicher Bewertung entsteht.
- Nachbearbeitung: Seit Prompt-Fassung 4.0 vergibt das Modell keine Kriterien-Scores mehr selbst. Für jedes Kriterium gibt es eine ZAEHLUNG-Zeile (die beobachtbaren Grössen) und eine BELEGE-Zeile (ein Zeitstempel pro gezählter Position) aus; die Auswertung wendet lokal die Skalenanker-Formel auf diese Zählungen an. Gesamtscore,

Spitzenwert, Schwere und Unser Mass werden ebenfalls lokal berechnet. Wenn die gezählte Auslösegrösse null beträgt, wird das Kriterium als «nicht anwendbar» statt als null ausgewiesen, und die Analyse muss dokumentieren, welche Kandidaten geprüft und verworfen wurden.

- Konvertierung: Berichte werden über einen zentralen Konverter (Fassung 4.2) mit einheitlicher Formatierung in standardisierte DOCX- und PDF-Formate umgewandelt. Bei Berichten der Prompt-Fassung 4.x liest der Konverter die berechnete Tabelle und rechnet nie nach; «nicht anwendbar» wird ohne Balken dargestellt, und Kriterien ohne Score werden als Anmerkung wiedergegeben. Ein Freigabe-Gate und ein Softfact-Sweep laufen vor der Konvertierung; nur Berichte, die beide bestehen, werden veröffentlicht.
- Aggregation: Einzelscores fliessen in länderspezifische und länderübergreifende Dashboards, Metaanalysen und Trendberichte ein.

Sämtliche arithmetischen Operationen (Mittelwertbildung, Trendberechnungen, statistische Tests) werden lokal durch das Auswertungsprogramm ausgeführt, nie durch das Sprachmodell. Diese Designentscheidung folgt aus dokumentierten systematischen Fehlern in der LLM-Arithmetik.

2.2 Modellauswahl

Die Produktion der Fassung 4 verwendet Claude Opus 4.6 (claude-opus-4-6) bei Temperatur 0; ältere Analysen stammen teils von Claude Sonnet 4.6. Werte verschiedener Modelle werden nicht gemischt, weil das Kriterienprofil vom Modell abhängt. Temperatur 0 beseitigt die Zufälligkeit des Samplings weitgehend, macht die Ausgabe aber nicht von Durchlauf zu Durchlauf identisch; das Verfahren ist nahezu deterministisch. Wie weit zwei Durchläufe derselben Sendung auseinanderliegen und wie genau Vergleiche zwischen Gruppen von Sendungen sind, ist in Abschnitt 8.3 beziffert. Aussagen über einen Sender über eine Reihe von Programmen beruhen auf der Gruppe.

2.3 Universelle Prompt-Architektur

Eine einzige Prompt-Vorlage dient allen Ländern. Länderspezifische Parameter werden zur Laufzeit eingesetzt:

- Sender und Sendegefass (etwa Hauptnachrichten, Debatte, Magazin)
- Anwendbarer Rechtsrahmen (Kurzverweis und vollständige Beschreibung)
- Politische Landschaft: Regierungskoalition, Opposition, Parteitabelle mit ideologischer Positionierung (CHES-Werte, sofern verfügbar), aktuelle Konfliktlinien
- Sprachanweisung (Analysesprache entspricht der Sendungssprache)
- Parteispezifische Bewertungstabelle (angepasst an das nationale Parteiensystem)

Analysestruktur, Kriteriendefinitionen, Bewertungsskala und Ausgabeformat sind über alle Länder hinweg identisch. Dies stellt sicher, dass ein Score von 5.0 in Norwegen dieselbe methodische Bedeutung trägt wie ein Score von 5.0 in Japan oder Argentinien. Formatregeln für maschinenlesbare Labels (Dimensionscodes, Quellenampel- Farben, Score-Überschriften) werden in jeder Sprache durchgesetzt — Labels werden nie übersetzt.

3. Analytischer Rahmen: Fünf Schritte

Jede Analyse folgt einer festen fünfstufigen Struktur. Die Reihenfolge spiegelt eine bewusste methodische Abfolge wider: zuerst der politische Kontext (um den Bezugsrahmen zu etablieren), dann die Inhaltsinventur, dann die kriterienweise Bewertung, dann die Synthese, dann die rechtliche Beurteilung.

Schritt 1 — Parteipolitische Bias-Bewertung

Für jede Partei im nationalen politischen System wird bewertet, ob die offiziellen Programmpositionen der Partei in der Sendung korrekt wiedergegeben wurden. Der Referenzmassstab ist das veröffentlichte Parteiprogramm, nicht die mediale Kommentierung über die Partei. Jede Partei erhält einen Score von -5 (Positionen aktiv verzerrt) über 0 (nicht erwähnt) bis +5 (Positionen korrekt und vollständig wiedergegeben). Ein Links-rechts-Tendenz-Score aggregiert den gesamten richtungsbezogenen Bias der Sendung. Diese politische Einordnung ist das 18. Kriterium (K19; K9 bleibt frei). Wo eine Partei auf der Links-Rechts-Achse steht, legt nicht das Modell fest, sondern der Chapel Hill Expert Survey (CHES 2024); für Länder, die CHES nicht abdeckt, gelten Expertenschätzungen.

Schritt 2 — Sendungsinformationen und thematischer Rahmen

Dieser Schritt inventarisiert die Sendung: Teilnehmende (mit Funktion, Zugehörigkeit, politischem Spektrum), Hauptthema, gesellschaftlicher Kontext und — entscheidend — eine Quellenampel-Bewertung für jeden Protagonisten (Abschnitt 5). Der Schritt identifiziert dann 8–10 relevante Perspektiven, die eine ausgewogene Sendung zu diesem Thema enthalten sollte (abgeleitet aus Allgemeinwissen, unabhängig vom Transkript), und gleicht jede mit dem tatsächlichen Sendungsinhalt ab. Fehlende Stimmen werden mit ihrem erwarteten Beitrag aufgeführt. Diese Inventur speist das Referenzraster von Kriterium K4, das das bewertete Instrument ist; Schritt 2 selbst erzeugt keinen Score.

Die kontextabhängige Ausgewogenheitsbewertung (Abschnitt 6) bestimmt den anwendbaren Ausgewogenheitsmassstab basierend auf dem Thementyp (wissenschaftlich, gesellschaftlich oder politisch).

Schritt 3 — 17 Qualitätskriterien: Detailanalyse

Der Kern der Analyse wendet die 17 Qualitätskriterien an (K1–K18, K9 aufgehoben), bei denen jeweils ausgewiesen ist, ob sie eindeutig zählbar oder auslegungsbedürftig sind (Abschnitt 4). Für jedes Kriterium identifiziert die Analyse bis zu 3 belegte Instanzen (die stärksten), jeweils mit: Zeitstempel, direktem Zitat, Beschreibung der Technik und Erklärung, warum sie problematisch ist. Die Bewertung leitet jeden Score aus den berichteten Zählungen auf einer Skala von 0 (neutral/ausgewogen) bis 10 (stark einseitig) ab; ein Kriterium, K15, wird ohne Score ausgewiesen.

Schritt 4 — Gesamtbewertung

Identifiziert die 3 dominanten Techniken, extrahiert die impliziten Botschaften der Sendung (Inhaltsebene, persönliche Ebene, gesellschaftliche Ebene) und ordnet den Befundgrad auf einer fünfstufigen Skala ein: ausgewogene Berichterstattung (0–2), leichte Tendenz (3–4), klare Einseitigkeit (5–6), systematische Schieflage (7–8), extreme Schieflage (9–10).

Schritt 5 — Rechtliche Beurteilung

Ordnet die identifizierten Befunde spezifischen Bestimmungen des anwendbaren nationalen Rundfunkrechts zu. Für jeden potenziellen Verstoss: die Norm, die tatsächliche Grundlage (mit Zeitstempel und Zitat) und die rechtliche Beurteilung. Dieser Schritt verankert die Analyse im regulatorischen Rahmen des jeweiligen Landes und ermöglicht es den Aufsichtsbehörden, die Befunde innerhalb ihrer Zuständigkeit zu bewerten.

4. Qualitätskriterien (K1–K18)

Das System misst 18 Kriterien: die 17 Qualitätskriterien dieses Abschnitts und als 18. die politische Einordnung der Parteien (K19; Schritt 1, Abschnitt 3). Kriterium K9 wurde in Fassung 4.0 aufgehoben; die Nummer bleibt unbesetzt, damit ältere Analysen lesbar bleiben. K15 wird beschrieben, trägt aber keinen Score; sechzehn Kriterien tragen einen Score. Die Quellenampel (Abschnitt 5) ist kein eigenes Kriterium, sondern ein Werkzeug der Expertenauswahl (K1). Alle bewerteten Kriterien beruhen auf gezählten Beobachtungen; sie unterscheiden sich darin, wie eng der Transkripttext die Zählung festlegt. Seit Fassung 4.14 trägt deshalb jedes Kriterium eine Einstufung: «eindeutig zählbar», wenn das Gezählte so im Text steht, dass man darauf zeigen kann, oder «auslegungsbedürftig», wenn vor dem Zählen zu entscheiden ist, was zählt. Grundlage ist die gemessene Abweichung zwischen zwei Durchläufen derselben Sendung (Abschnitt 8.3; die Zahl in Klammern bei jedem Kriterium ist die mittlere Abweichung in Punkten, der Schnitt liegt bei rund 1.3), wo Daten fehlen die Bauart des Kriteriums.

Die frühere Einteilung in Hardfacts (K1–K8) und Softfacts (K10–K18) hielt dieser Messung nicht stand: Vollständigkeit (K4, 1.64) und Timing (K7, 1.72) streuen stärker als der Schnitt, die Vorverurteilung (K17, 0.36) ist das stabilste Kriterium. Die Einstufung ist methodisch bedeutsam, weil sie zeigt, wo sich über eine Zählung streiten lässt; sie kann je nach Auftrag im Bericht ausgewiesen oder weggelassen werden.

Epistemologischer Status der analytischen Kategorien. Die 18 Kriterien und das Quellenampel-Rahmenwerk leiten sich aus den operativen Anforderungen nationaler Rundfunkgesetze ab, nicht aus normativen Prämissen der Autoren. Im Kodierschema verwendete Begriffe — wie «qualified dissident» (jemand, dessen Position der erwarteten Haltung seiner institutionellen Zugehörigkeit widerspricht, dokumentiert durch überprüfbare berufliche Konsequenzen) oder «silent majority» (eine messbare Diskrepanz zwischen Umfragedaten und medialer Darstellung) — beschreiben empirisch beobachtbare, operational definierte Muster und keine Werturteile. Das Rahmenwerk ist rechtspositivistisch angelegt: Es operationalisiert, was die Rundfunkgesetzgebung verlangt, und beurteilt die Einhaltung anhand dieser Anforderungen. Wo Kriterien Auslegung erfordern, ist dies beim Kriterium ausgewiesen (auslegungsbedürftig), und jeder Befund muss mit textlichem Beleg aus dem Transkript dokumentiert werden. Das Kodierschema setzt nicht voraus, dass ein bestimmtes Berichterstattungsergebnis wünschenswert ist; es misst die Abweichung vom gesetzlichen Standard ausgewogener, pluralistischer und korrekter Berichterstattung, wie er im Rundfunkrecht des jeweiligen Landes definiert ist.

4.1 Kriterien K1 bis K8

K1 — Expertenauswahl. Wer tritt als Experte auf? Bewertet Vielfalt, Unabhängigkeit und Qualifikation der Expertenstimmen. Umfasst eine vertiefte Quellenprüfung (Finanzierung, Mandat, Interessenkonflikte) und die Quellenampel-Bewertung für jeden Experten. Mediale Rahmung von Experten («renommiert», «umstritten») wird als analytischer Befund behandelt, nicht als Tatsache. Gezählt werden: die Anzahl der umstrittenen Fachfragen, die dazu auftretenden Expertenstimmen, wie viele davon die dominante Position vertreten und wie viele ein nicht offengelegtes Interesse haben. Wenn die Sendung keine umstrittene Fachfrage aufwirft — ein Spielkommentar, ein Porträt, eine Ergebnissendung —, ist das Kriterium nicht anwendbar, und die Analyse vermerkt, welche Kandidaten geprüft wurden. Einstufung: eindeutig zählbar (1.18).

K2 — Quellenauswahl. Welche Quellen werden zitiert? Bewertet Vielfalt, Unabhängigkeit und strukturelle Interessenkonflikte der zitierten Quellen (NGOs, Regierungsstellen, kirchlich finanzierte Einrichtungen, Interessenorganisationen). Gezählt werden die Quellen von Tatsachenbehauptungen, solche ohne benennbaren Träger («angeblich», «aus Kreisen verlautet», «Berichten zufolge») und solche aus dem dominanten Lager des Beitrags; der Score ergibt sich aus den beiden Anteilen. Einstufung: eindeutig zählbar (0.65).

K3 — Zeitverteilung. Redezeit-Zuteilung zwischen den Positionen. Sendungsdauer, Sprechzeit und Wortzahl werden vor Beginn der Analyse programmatisch aus den Untertitel-Zeitangaben

gemessen und als bindende Tatsachen in den Prompt injiziert; das Modell darf sie nicht überschreiben. Die Zuteilung auf die Positionen schätzt das Modell, weil die Untertitel keine Sprecher ausweisen. Einstufung: eindeutig zählbar — Sendungsdauer und Sprechzeit werden programmatisch aus den Untertiteln gemessen (0.41).

- K4 — Vollständigkeit.** Was wird trotz Relevanz nicht gezeigt? Seit Fassung 4.0 wendet das Kriterium ein Referenzraster aus acht Zellen an (vier Rahmungsfunktionen nach Betroffenen und Entscheiden- den); die Analyse bestimmt, welche Zellen für das Thema relevant sind und welche davon die Sendung unberührt lässt. Eine separate Zählung erfasst den Fall, den die schweizerische Aufsichtsbehörde am häufigsten gutheisst: einen Beitrag, in dem eine substantiierte Stellungnahme der kritisierten Partei vorlag und nur deren Ergebnis berichtet wurde. Eine blossе Bestreitung genügt, sofern sie berichtet wird (Entscheid b.617). Einstufung: auslegungsbedürftig — welche Zellen des Rasters relevant sind, ist ein Urteil (1.64).
- K5 — Zahlendarstellung.** Selektiver oder irreführender Umgang mit Zahlen. Nur Zahlen mit argumentativer Funktion werden gezählt. Für jede Zahl bestimmt die Analyse zunächst, welche von drei Dimensionen anwendbar sind — absoluter Wert, Anteil/Pro-Kopf, Trend über die Zeit — und dann, wie viele davon fehlen; irreführend dargestellte Zahlen (Kategorien zusammengelegt, eine Schätzung als Tatsache dargestellt, Ungleiches als Gleiches verglichen) setzen das Kriterium auf sein Maximum. Die Zuordnung einer Zahl zu einer Quelle wird nicht mehr hier gezählt, sondern unter K18. In den Berichten der Fassung 4.14 trägt das Kriterium noch den früheren Namen «Zahlen-Manipulation». Einstufung: eindeutig zählbar (1.21).
- K6 — Guilt by Association.** Diskreditierung durch Assoziation mit negativen Gruppen oder Ideen. Umfasst eine strukturierte Gegenprüfung für alle, die als «Verschwörungstheoretiker» oder ähnlich bezeichnet werden: Primärquellen? Falsifizierbare Behauptungen? Risikomatrix (was hat die bezeichnete Quelle verloren, was gewonnen?). Kategorisierung: A (Forscher mit Methodik und Belegen), B (Grenzfall: teils belegt, teils spekulativ), C (Ideologe: unbelegte, nicht falsifizierbare Behauptungen). Mediale Rahmung ist kein Beleg für die Kategorisierung. Wird symmetrisch in beide Richtungen angewandt. Einstufung: auslegungsbedürftig nach Bauart — ob ein Sachzusammenhang dargestellt ist, ist ein Urteil.
- K7 — Timing.** Platzierung belastender und entlastender Elemente innerhalb der Sendung. Gezählt werden alle Elemente, die sich auf denselben Vorwurf beziehen, und darunter diejenigen, die prominent platziert sind (Titel, Anmoderation, Eröffnung des Beitrags); der Score ist die Differenz zwischen den prominent platzierten belastenden und entlastenden Elementen, relativ zu allen. Einstufung: auslegungsbedürftig — was als hervorgehoben gilt, ist ein Urteil (1.72).
- K8 — Selektive Empörung.** Empörung über bestimmte Positionen, aber nicht über vergleichbare andere. Methodisches Prinzip: Das auslösende Ereignis muss dokumentiert sein, bevor eine Asymmetrie beurteilt werden kann. Fälle werden nach Schwere gezählt und nicht als Anteil: Ein schwerer Fall erfordert, dass eine Seite einen Empörungsmarker trägt, während der vergleichbare Fall im selben Beitrag keinen trägt, und dass kein sachlicher Unterschied, der die Asymmetrie rechtfertigt, genannt wird. Ohne vergleichbaren Fall ist das Kriterium nicht anwendbar. Einstufung: auslegungsbedürftig nach Bauart — ob zwei Vorgänge gleichartig sind, ist ein Urteil.
- K9 — Aufgehoben (Fassung 4.0).** Gemessen über 1'393 Analysen der 3.x-Reihe korrelierte K9 mit K4 bei $r = 0.987$ und trug in 90 Prozent der Berichte einen identischen Score: Zwei Kriterien, die dasselbe messen, gingen doppelt in den Gesamtscore ein. Sein Gegenstand wird nun durch das Referenzraster in K4 abgedeckt. Die Nummer bleibt unbesetzt.

4.2 Kriterien K10 bis K18

- K10 — Framing.** Zwei Lesarten, nebeneinander ausgewiesen. Die UBI-Messart fragt, was die schweizerische Aufsichtsbehörde fragt: Wurde die Perspektive angekündigt, und hält die Anmoderation dem Beitrag stand? Gemessen an 203 dokumentierten Fällen hiess die UBI eine Beschwerde gegen eine angekündigte Perspektive in 5 Prozent der Fälle gut, gegen eine nicht angekündigte, die einen falschen Gesamteindruck erzeugt, in 63 Prozent; Ofcom Rule 5.10 ist inhaltlich identisch. Unser Mass, im Bericht zuerst genannt, stellt eine andere Frage: In wie vielen Beiträgen erscheint nur eine Seite der strittigen Sache, und in wie vielen davon wird die Gegenposition nicht einmal benannt. Zusätzlich und ohne Score hält der Bericht in Prosa fest, wie die Sendung ihren Gegenstand erzählt — mit Zitat und Zeitstempel — und was ein Gegengewicht gekostet hätte: einen O-Ton, einen Satz Kontext, eine Nachfrage. Einstufung: auslegungsbedürftig (2.57).
- K11 — Wortwahl und Begriffe.** Welche Sprache wird verwendet und welche Konnotationen werden gesetzt? Zu jedem Befund wird die neutrale Alternative explizit genannt. Gezählt werden die wertenden Begriffe und die Seite, auf die sie jeweils zielen — nicht wer sie äussert. Unter fünf solchen Begriffen wird kein Verhältnis gebildet: Drei auf eine Seite zielende Begriffe sind ein Befund, aber kein Mass, und würden sonst dasselbe Maximum tragen wie dreissig. Einstufung: auslegungsbedürftig — ob ein Begriff wertend ist, ist ein Urteil (2.49).
- K12 — Moderationsverhalten.** Asymmetrien bei Nachfragen, Unterbrechungen und Sympathiebekundungen. Gezählt wird pro Interviewpartner: gegebene Antworten und mitten im Satz abgebrochene Antworten. Welcher von zwei Interviewpartnern häufiger unterbrochen wurde, ist eine Berechnung, keine Beobachtung, und wird durch die Auswertung bestimmt, nicht durch das Modell. Methodischer Grundsatz: Das auslösende Ereignis muss dokumentiert sein. Ein Eingriff kann nur dann als asymmetrisch bewertet werden, wenn vergleichbare Auslöser bei anderen Gästen keinen analogen Eingriff hervorriefen. Einstufung: auslegungsbedürftig — in 22 Prozent der Doppelmessungen wechselt die Anwendbarkeit.
- K13 — Fragen-Asymmetrie.** Unterschiedliche Härte der an verschiedene Teilnehmende gerichteten Fragen. Gezählt werden pro Interviewpartner die gestellten Fragen und darunter jene, die einen Vorwurf oder eine Gegenbehauptung enthalten; der Score ist die Differenz zwischen dem höchsten und dem niedrigsten Anteil. Bei nur einem Interviewpartner ist das Kriterium nicht anwendbar — dass die Gegenseite fehlt, wird durch K1 und K3 erfasst. Einstufung: eindeutig zählbar — ein Vorhalt ist am Wortlaut erkennbar (0.82 an vier Paaren).
- K14 — False Balance.** Künstliche Ausgewogenheit trotz tatsächlichem Ungleichgewicht: Eine belegbare und eine unbelegte Position stehen wie gleichwertig nebeneinander, mit vergleichbarem Raum, und kein Satz im Beitrag ordnet den Unterschied ein. Massstab ist die Belegbarkeit der Aussagen, nicht die Glaubwürdigkeit der Sprechenden; diese bewertet die Quellenampel bei K1. Fälle werden nach Schwere gezählt und nicht als Anteil: In realen Sendungen ist der Nenner fast immer eins, und ein Anteil über einen Fall sagt nichts aus. Einstufung: auslegungsbedürftig nach Bauart — die Belegbarkeit einer Position ist einzuschätzen.
- K15 — Agenda-Setting.** Was gilt als normal/selbstverständlich? Was schafft es nicht auf die Agenda? Das Kriterium trägt keinen Score in der Ergebnistabelle: Über 865 Analysen der 3.x-Reihe lag es im Durchschnitt bei 4.03, war aber nur in 6 Prozent der Sendungen der höchste Wert, und die schweizerische Aufsichtsbehörde behandelte Agenda-Setting in 2 von 282 untersuchten Geschäften. Zählbar ist die ungeprüfte Prämisse — die Vorfrage, die die Sendung nie stellt — und diese Zählung erscheint nur in der Spalte Unser Mass. Einstufung: auslegungsbedürftig nach Bauart — die unbefragte Prämisse ist eine Einschätzung.
- K16 — Behutsamkeit und Menschenwürde.** Wird ein identifizierbarer Mensch als Mensch herabgesetzt statt für sein Verhalten kritisiert, oder in einem Zustand der Not gegen seinen erkennbaren Willen gezeigt? Fälle werden nach Schwere gezählt; drei oder mehr identifizierende Merkmale zusammen mit besonderer Verletzlichkeit (Trauer, Krankheit, Kinder) bilden den schweren Fall. Eingeführt in Fassung 4.0, nachdem 13 von 130 untersuchten Aufsichtsentscheiden Sachverhalte betrafen, für die kein Kriterium existierte — neun davon Menschenwürde, vier die

Unschuldsvermutung, die nun K17 ist. Vorbild ist die schwedische Granskningsnämnden, die «varsamhet» — Behutsamkeit gegenüber Betroffenen — als eigenen Prüfpunkt führt, unabhängig davon, ob ein Verfahren läuft; in 17 von 385 ausgewerteten Fällen zu SVT, SR und UR war sie Prüfgegenstand. Das Kriterium übernimmt diese Trennung: Wer herabgesetzt oder blossgestellt wird, ist ein eigener Befund, auch ohne Vorwurf und ohne Verfahren. Einstufung: auslegungsbedürftig nach Bauart — die Herabsetzung eines Menschen statt Kritik an seinem Handeln ist ein Urteil.

K17 — Vorverurteilung. Wird jemand, gegen den ein Verfahren hängig ist, als verurteilt dargestellt? Gezählt werden die Betroffenen, die Fälle, in denen Schuld ohne Vorbehalt behauptet wird, und jene, in denen sie lediglich impliziert wird. Ob ein Verfahren hängig ist, wird der Sendung selbst entnommen, nicht einer Recherche: Die Aufsichtsbehörde beurteilt die Sachlage ebenfalls so, wie sie zum Zeitpunkt der Ausstrahlung bestand. Einstufung: eindeutig zählbar — das stabilste Kriterium (0.36; in 91 Prozent der Doppelmessungen identisch).

K18 — Quellenbelegung. Werden zentrale Behauptungen einer namentlich benennbaren, überprüfbaren Quelle zugeordnet, und werden die Interessen der Auftretenden offengelegt? Gezählt werden die zentralen Behauptungen und jene ohne Attribution («Studien zeigen», «Experten sind sich einig»). Quellenbelegung ist ein eigenständiger Prüfpunkt in sieben der elf Einzelfälle für Einzelfall untersuchten Aufsichtskorpora. Einstufung: auslegungsbedürftig — ob eine Quelle nachprüfbar benannt ist, ist teils ein Urteil (1.48).

5. Quellenampel

Die Quellenampel ist keine eigene Prüfmethode, sondern ein Werkzeug der Expertenauswahl (K1). Jeder in einer Sendung auftretende Protagonist — Fachleute, Interviewpartner, zitierte Spezialisten, Vertreter — wird auf sechs Dimensionen bewertet, die jeweils von -2 bis +2 skaliert sind. Die Bewertung basiert auf dem allgemeinen Wissen über die Quelle, NICHT darauf, wie die Medien sie rahmen. Die mediale Rahmung («renommierter Experte», «führender Forscher», «umstritten», «Verschwörungstheoretiker») ist selbst ein analytischer Befund, keine Tatsache.

D1 — Interessenkonflikt. Aufgeteilt in D1a (direktes finanzielles, institutionelles oder ideologisches Eigeninteresse) und D1b (strukturelle Patronageabhängigkeit). D1b verwendet 5 Indikatoren:

- (1) finanzielle Verstärkung durch einen dominanten Geldgeber,
- (2) Netzwerk-Kooptation,
- (3) Karriere-Gatekeeper-Abhängigkeit,
- (4) einseitige Unterstützungskette,
- (5) institutionelle Monokultur (>50% aus einem Ökosystem).

0-1 Indikatoren = 0, 1-2 = ± 1 , 3+ = ± 2 .

D1 = max(|D1a|, |D1b|) mit Vorzeichen. Symmetrisch über alle politischen Richtungen angewandt.

D2 — Persönliches Risiko. Was riskiert die Quelle mit ihrer Position? Karriereverlust, soziale Ausgrenzung, finanzielle Konsequenzen. Hohes persönliches Risiko erhöht die Glaubwürdigkeit.

D3 — Fachkompetenz. Spricht die Quelle innerhalb oder ausserhalb ihres Fachgebiets? Primärforschende innerhalb ihres Gebiets erhalten die höchste Bewertung.

D4 — Meinungskonsistenz. Unterscheidet drei Arten von Positionswechseln: (a) Opportunistisches Mitläufertum mit dem politischen Wind (negativ), (b) oberflächliche, aber konsistente Positionierung (neutral), (c) selbstkritische Korrektur gegen die eigene frühere Position mit Begründung (positiv). Selbstkorrektur wird angerechnet; korrigiert zu werden und dann mitzuziehen nicht.

D5 — Emotionalisierung vs. Daten. Ist die Argumentation datenbasiert oder appellativ? Evidenzgetriebener Diskurs wird positiv bewertet; emotionale Appelle ohne Untermauerung werden negativ bewertet.

D6 — Quellenstufe. Primärquelle (+2) / Sekundärquelle (0) / Tertiärquelle (-2). Misst die Nähe zu den Originaldaten oder dem Ereignis.

Die sechs Dimensions -Scores werden zu einem TOTAL (-12 bis +12) summiert. Die Ampelklassifikation folgt: GRÜN ($\geq +5$), GELB (-4 bis +4), ROT (≤ -5). Jede Quelle wird zusätzlich kategorisiert: A (Forscher mit Methodik und Primärdaten), B (Grenzfall), C (Advocacy: Standpunktvertretung ohne eigene Primärdaten), N (nicht einordenbar: Politik/Aktivismus), 0 (keine belegbare Begründung).

Entscheidend ist, dass Diskrepanzen zwischen dem Ampelergebnis und der medialen Rahmung Befunde sind. Eine als «renommierter Experte» gerahmte Quelle kann GELB oder ROT bewertet werden (z. B. ein CEO mit Interessenkonflikt). Eine als «umstritten» gerahmte Quelle kann GRÜN bewertet werden (z. B. ein systemkritischer Forscher mit Methodik, persönlichem Risiko und Primärdaten). In beiden Fällen fließt die Diskrepanz in K1, K6, K10 oder K11 ein.

Die Quellenampel wurde über 27 thematische Matrizen gegeneinander geprüft, die die wichtigsten politischen Themen in der Schweiz abdecken, dabei wurden 219 Einstufungen präzisiert und 329 ergänzt (Schlaepfer 2026c). Die Dimension D3 (Fachkompetenz) wurde an einem äusseren Massstab geprüft: Bei 453 Fachleuten aus 25 umstrittenen Themen hängt die Einstufung mit der wissenschaftlichen Publikationstätigkeit zusammen (Spearman $\rho = 0.591$, $n = 453$; Schlaepfer 2026d).

Interkulturelle Operationalisierung. Jede der sechs Quellenampel-Dimensionen wird durch beobachtbare, überprüfbare Indikatoren operationalisiert, die über politische Systeme und Medienumgebungen hinweg anwendbar sind.

D1 (Interessenkonflikt) stützt sich auf öffentlich überprüfbare Fakten: institutionelle Anstellung, Verwaltungsratsmandate, Finanzierungsbeziehungen und organisatorische Zugehörigkeiten.

D1b (Strukturelle Patronage) wird anhand von fünf Indikatoren bewertet: Karriereabhängigkeit von einer einzelnen Institution, Gatekeeping-Macht dieser Institution, Fehlen alternativer Karrierewege, öffentlich dokumentierte Loyalitätserwartungen und Muster der Positionsübereinstimmung mit institutionellen Interessen.

D2 (Persönliches Risiko) bewertet dokumentierte oder plausible Konsequenzen einer Positionierung: Karrieregefährdung, soziale Ächtung, finanzieller Verlust oder physische Bedrohung, kontextualisiert im spezifischen politischen Umfeld.

D3 (Fachliche Kompetenz) wird anhand überprüfbarer Qualifikationen bewertet: akademische Abschlüsse, Publikationsverzeichnisse und ob die Quelle innerhalb ihres dokumentierten Fachgebiets spricht.

D4 (Meinungskonsistenz) erfordert eine Längsschnittbeobachtung öffentlich dokumentierter Positionen.

D5 (Emotionalisierung vs. Daten) unterscheidet evidenzbasierte Argumentation von appellativer Rhetorik durch textuelle Analyse auf Transkriptebene.

D6 (Quellenstufe) kategorisiert die Nähe zu Originaldaten oder Ereignissen auf einer dreistufigen Skala. Eine unabhängige Replikation erfordert Zugang zu denselben öffentlichen biografischen und institutionellen Informationen, die bei der ursprünglichen Bewertung verwendet wurden; ein Replikationsprotokoll, das die erforderlichen Datenquellen für jede Dimension spezifiziert, ist auf Anfrage bei den Autoren erhältlich.

6. Kontextabhängige Ausgewogenheitsbewertung

Nicht alle Themen erfordern denselben Ausgewogenheitsstandard. Das System unterscheidet drei Thementypen, jeweils mit einem eigenen Bewertungsrahmen:

6.1 Wissenschaftliche Themen

Für Themen, die Forschung, Technologie, Medizin oder Klima betreffen: Konsens wird als Arbeitsinstrument behandelt, nicht als Wahrheitsbeweis. Besondere Aufmerksamkeit gilt qualifizierten Dissidenten, die an der Grenze des etablierten Wissens arbeiten. Die zentrale Unterscheidung: qualifizierter Dissens (Primärdaten, peer-reviewed, persönliches Risiko) ist NICHT dasselbe wie unbegründete Gegenbehauptungen. Die systematische Ausschliessung von Forschenden der Kategorie A (Quellenampel), die Minderheitspositionen vertreten, ist ein Befund, kein Ausgewogenheitsschutz.

6.2 Gesellschaftliche Themen

Für Themen, die Wohnen, Alter, Geschlecht, Sozialpolitik, Gesundheit betreffen: alle betroffenen Gruppen müssen vertreten sein (Konsensregel). Dies schliesst ausdrücklich diejenigen ein, die das System finanzieren (Steuerzahlende, Prämienzahlende, Beitragszahlende).

Ausnahme: Wer nicht dauerhaft betroffen ist (z. B. Gäste in einem Land), hat keinen Anspruch auf Mitbestimmung bei Infrastruktur, die er nicht finanziert. Die Geschlechterabdeckung erfordert beide Seiten; wo Gesetze oder strukturelle Ungleichgewichte systematisch eine Seite begünstigen, wird dies als Befund markiert. Politische Korrektheit ist kein Massstab; Verhältnismässigkeit schon.

6.3 Politische Themen

Für Themen, die Parteien, Wahlen, Gesetzgebung, Abstimmungen betreffen: Parlamentsstärke und Volksmehrheiten dienen als Massstab für die Redezeitverteilung. Parteiscores aus Schritt 1 liefern die Referenz. Ein kritischer methodischer Punkt: die schweigende Mehrheit — jene, die nicht zu Wort kommen — ist in der Sendungsberichterstattung systematisch untervertreten. Bei Abstimmungen dient das tatsächliche Ergebnis als härtester Massstab.

7. Scoring-Methodik

7.1 Score -Skala

Alle Kriterien werden auf einer Skala von 0–10 bewertet, wobei 0 für neutrale, ausgewogene Berichterstattung steht und 10 für stark einseitige Inhalte (dies bezeichnet eine messbare Wirkung, nicht eine redaktionelle Absicht). Diese Richtung ist wesentlich: ein hoher Score zeigt ein Problem an, nicht Qualität. Die Skala ist über alle Kriterien und alle Länder hinweg einheitlich.

7.2 Gesamtscores

Die folgenden Kennzahlen werden lokal aus den ZAEHLUNG-Zeilen berechnet (nie durch das Sprachmodell):

- **Gesamtscore** = arithmetisches Mittel aller anwendbaren bewerteten Kriterien; der Bericht nennt die Basis (etwa «Mittel über 14 von 16 bewerteten Kriterien»)
 - **Spitzenwert** = der einzelne schwerwiegendste Befund, mit Nennung des Kriteriums
 - **Schwere** = Mittelwert der drei höchsten Werte unter den neun Kriterien, die dem entsprechen, was die Aufsichtsbehörde prüft (K2, K3, K4, K5, K10, K11, K12, K16, K17)
 - **Unser Mass** = Mittelwert der drei höchsten Werte unter den zehn Kriterien, die dieses Projekt als notwendig erachtet (K1, K5, K6, K7, K8, K10, K13, K14, K15, K18), unter Verwendung der eigenen Lesart, wo zwei existieren; wird vor dem Schwere-Wert ausgewiesen und als Position gekennzeichnet, nicht als Feststellung einer Rechtswidrigkeit
- Kriterien, deren Auslösemenge null beträgt, werden aus den Mittelwerten ausgeschlossen, anstatt als null

gezählt zu werden. Ein Gesamtscore entfällt nur, wenn kein einziges Kriterium anwendbar ist; ist mehr als ein Drittel nicht anwendbar, ist er nur eingeschränkt vergleichbar.

Einen Hardfacts- und einen Softfacts-Score gibt es seit Fassung 4.14 nicht mehr (Abschnitt 4). Ihr Wegfall verschob an 265 Berichten der Fassung 4 den Gesamtscore im Mittel um 0.17 Punkte; die Rangfolge der Sendungen blieb erhalten (Spearman 0.977). Das einfache Mittel ist ruhiger als das frühere Mittel zweier Gruppenmittel: Zwei Messungen derselben Sendung liegen im Gesamtscore im Mittel 0.71 Punkte auseinander statt 0.82 (Abschnitt 8.3). Wer Kriterien anders gewichten will, findet in der Tabelle jedes Berichts die Einzelwerte und ihre Einstufung.

7.3 Vom Zählwert zum Score

Bei jedem Kriterium mit Zählregel vergibt nicht das Sprachmodell den Score, sondern die Auswertung. Das Modell liefert dazu je Kriterium die Zählwerte der vorgegebenen Grössen (Zeile ZAEHLUNG), zu jeder gezählten Position die Zeitmarke in der Sendung (BELEGE) und, wo ein Kriterium nicht anwendbar ist, den Beleg der Suche (GEPRUEFT). Ein Programm liest diese Zeilen und rechnet den Score nach der festen Regel des Kriteriums aus, für alle Sendungen gleich. Nennt das Modell beim Kriterium selbst einen Wert, gilt der gerechnete; weichen die beiden um einen halben Punkt oder mehr ab, hält das Laufprotokoll beide fest. Eine Zählregel tragen 15 der 16 bewerteten Kriterien. Bei K3 misst das Programm Sendungsdauer, Sprechzeit und Wortzahl (Abschnitt 4); die Zuteilung der Redezeit schätzt das Modell, und den Wert vergibt es selbst. Wo ein Kriterium zwei Lesarten kennt — die der Aufsicht und die eigene (Abschnitt 7.2) —, zählt das Modell beide, und das Programm rechnet beide.

Bevor es rechnet, prüft das Programm die Zählungen. Für jede gezählte Grösse muss die Zeile BELEGE so viele Zeitmarken nennen, wie gezählt wurden; sonst meldet es die Zählung als unbelegt. Fehlt die Zählzeile oder ist sie unbrauchbar — es fehlen Grössen, ein Wert ist negativ, oder die Regel ergibt einen Wert ausserhalb von 0 bis 10 —, bleibt der Wert des Modells stehen und wird als solcher gemeldet. Ergibt die Regel «nicht anwendbar», weil die Bezugsgrösse in der Sendung nicht oder zu selten vorkommt, fällt das Kriterium aus dem Mittel, statt als null gezählt zu werden; fehlt dazu der Beleg der Suche, wird auch das als unbelegt gemeldet. Fehlt für ein Kriterium jeder Wert, bildet das Programm keinen Gesamtscore. Die Zeilen werden über die Nummer des Kriteriums und das Format der Zeitmarken gelesen und sind damit von der Sprache des Berichts unabhängig.

8. Validierungsrahmen

Die Messqualität wird in dreizehn Abschnitten geprüft, von denen jeder einen eigenständigen Aspekt adressiert. Jede nachstehende empirische Zahl nennt die Stichprobe, auf der sie beruht, und die Prompt-Version, unter der sie gemessen wurde; Zahlen aus verschiedenen Versionen werden nicht zusammengefasst.

8.1 Spezifität (Sport-Kalibrierung)

Sportsendungen dienen als Kontrollgruppe. Im Oktober 2026 wurden mit der Fassung 4.14 zehn ganze Spielübertragungen der SRG gemessen: drei aus dem Jahr 2026 — das Eröffnungsspiel der Fussball-WM (RTS), ein Nations-League-Spiel der Nationalmannschaft und ein Champions-League-Spiel (RSI, je mit Studio vor und nach dem Spiel) — und sieben 2020 von SRF wiederholte Spiele mit Schweizer Beteiligung aus den Jahren 2009 bis 2018: fünf Eishockeyspiele, davon vier der Nationalmannschaft, und zwei Fussballspiele der Super League. Der Gesamtscore lag zwischen 0.0 und 1.8 (Mittel 0.6, Median 0.1), die Schwere, die sich bei sieben Spielen bilden liess, zwischen 0.0 und 3.4. Anwendbar waren je eines bis zehn der sechzehn Kriterien; nach der Regel der Fassung 4.14 sind alle Werte damit nur eingeschränkt vergleichbar — das Fehlen von Messbarem wird nicht als Befund verbucht. Zum Vergleich: Sendungen, deren Beschwerde die UBI abwies, liegen je nach Format im Mittel zwischen 3.0 und 5.7 im Gesamtscore und zwischen 4.4 und 8.1 in der Schwere (Formate mit mindestens vier abgewiesenen Sendungen).

Ein niedriger Score bei einem Spielkommentar belegt nicht Ausgewogenheit; er zeigt, dass das Rahmenwerk wenig findet, wofür es gebaut ist. Wo eine Schweizer Mannschaft spielt, nimmt der Kommentar ihre

Sicht ein, und das Instrument zählt es: Im Viertelfinal der Eishockey-WM 2013 gehörten alle sieben Stimmen, die zu Wort kamen, zum Schweizer Team (Quellenauswahl 5); beim Meisterspiel der Young Boys 2018 verteilte sich die Redezeit etwa zwei zu eins zugunsten der Berner Sicht (Zeitverteilung 3). Beim Nations-League-Spiel und beim WM-Viertelfinal setzt die eigene Lesart das Framing (K10) auf den Höchstwert 10, die UBI-Messart auf 0, weil der Schweizer Blickwinkel deklariert ist.

8.2 Diskriminante Validität (Ländervergleich)

Wenn das System reale Unterschiede in der redaktionellen Praxis misst, sollten die Scores systematisch zwischen Ländern mit bekannten Unterschieden in der Pressefreiheit und redaktionellen Unabhängigkeit variieren. Ein Ländervergleich ist nur innerhalb einer Versionsfamilie und eines Programmformats aussagekräftig, da beide die Skala verschieben; Vergleiche, die diese beiden Bedingungen erfüllen, werden in den Länderstudien berichtet und nicht hier.

8.3 Intra-Rater-Reliabilität (Wiederholungsmessung)

Dasselbe Transkript, mehrfach analysiert, sollte konsistente Scores ergeben. Unter der Fassung 4 (Unterfassung 4.9) wurden 23 Sendungen (12 mit gutgeheissener, 11 mit abgewiesener Beschwerde) zweimal unter identischen Bedingungen analysiert. Zwei Durchläufe weichen im Mittel um 0.71 Punkte im Gesamtscore (Rechenart der Fassung 4.14) und um 1.29 Punkte in der Schwere voneinander ab; die Rangfolge der Sendungen bleibt teilweise erhalten (Spearman zwischen den Durchläufen +0.51 bzw. +0.56). Die Abweichung ist je Kriterium verschieden: am kleinsten bei der Vorverurteilung (K17: 0.36, in 91 Prozent der Paare identisch) und bei der programmatisch gemessenen Zeitverteilung (K3: 0.41), am grössten beim Framing (K10: 2.57) und bei der Wortwahl (K11: 2.49). Ein Einzelwert ist damit ein Wert mit Band; Unterschiede unter einem Punkt sind keine Unterschiede. Für Aussagen über Gruppen von Sendungen genügt die Genauigkeit eher; dort zählt aber vor allem, wie stark die Sendungen selbst streuen. Rechnet man diese Streuung mit ein, beträgt der Standardfehler des Unterschieds zwischen gutgeheissenen und abgewiesenen Sendungen im selben Format rund 0.15 Punkte im Gesamtscore und 0.26 in der Schwere (Abschnitt 8.12).

Jeder Schweizer Bericht nennt diese Streuung und weist aus, ob er eine Einzelmessung oder das Mittel aus zwei Messungen ist; belastbar für eine einzelne Sendung — insbesondere bei forensischem oder rechtlichem Gebrauch — ist das Mittel aus zwei Durchläufen. Für Aussagen über einen Sender über eine Reihe von Programmen hinweg ist eine wiederholte Erstellung nicht nötig: Ab ungefähr 36 Sendungen fällt das Rauschen des Gruppenmittelwerts unter 0.2 Punkte.

8.4 Reproduzierbarkeit

Deterministisch ist die Berechnung. Jeder Score folgt aus den ZAEHLUNG-Zeilen nach einer veröffentlichten Formel; bei gleichen Zählungen ergeben sich die gleichen Zahlen, und jede Drittpartei kann sie anhand des Berichts selbst nachrechnen. Jede gezählte Position trägt einen Zeitstempel, sodass die Zählungen am Transkript überprüft werden können. Temperatur 0 beseitigt die Stichprobenzufälligkeit aus dem Leseschritt; die verbleibende Variation zwischen zwei Lesungen desselben Transkripts wird in Abschnitt 8.3 quantifiziert und bei jeder Aussage über eine einzelne Sendung ausgewiesen.

8.5 Modellvergleich

Für die Fassung 4 liegt kein Vergleich mehrerer Modelle vor. Unter der Fassung 3 wurde eine geschichtete Stichprobe von 98 Schweizer Sendungen von drei Modellen verschiedener Anbieter gemessen (Claude Sonnet 4.6, Gemini 2.5 Flash, GPT-4o mini): Alle drei ordneten die Sendungen im Mittel derselben Richtung zu; auf der Ebene einzelner Werte stimmten sie deutlich weniger überein (Schlaepfer 2026a). Werte der Fassungen 3 und 4 sind nicht vergleichbar: An 19 Sendungen, die unter beiden Fassungen gemessen wurden, blieb die Rangfolge kaum erhalten (Spearman +0.25). Die Produktion verwendet ein einziges Modell; Werte verschiedener Modelle werden nicht gemischt (Abschnitt 2.2).

8.6 Rechnerische Korrektheit

Sämtliche Berechnungen werden lokal ausgeführt, nicht durch das Sprachmodell. Die aktuelle Pipeline erzwingt eine rechnerische Trennung: Das Modell liefert Zählungen und zeitgestempelte Belege, die lokale Auswertung berechnet jeden Kriterienwert und jedes Aggregat. Das Modell wird angewiesen, den Ergebnisblock überhaupt nicht zu schreiben.

8.7 Konstruktvalidität (Quellenampel — 27 Matrizen)

Unter der 3.x-Reihe wurde die Quellenampel auf 27 thematische Policy-Matrizen in der Schweiz angewandt, die jeweils die wichtigsten Protagonisten einer politischen Debatte abdecken (von Abtreibung bis Kernenergie). Anschliessend wurden alle 27 Matrizen systematisch gegeneinander geprüft: Wer in mehreren Themenfeldern auftritt, muss dort nach denselben Massstäben eingestuft sein. Diese Qualitätssicherung hat 219 Einstufungen präzisiert und 329 ergänzt; seither sind die Bewertungen über die Politikbereiche hinweg einheitlich.

8.8 Universalität (ein Prompt, alle Länder)

Eine einzige Prompt-Vorlage dient allen Ländern. Länderspezifische Rechtsrahmen, politische Landschaften und Parteiensysteme werden als Parameter eingefügt. Die analytische Struktur, die Kriterien und die Bewertung bleiben identisch. Dieses Design stellt sicher, dass länderübergreifende Vergleiche tatsächliche Unterschiede in der redaktionellen Praxis widerspiegeln und nicht methodische Variation.

8.9 Verhältnis zur klassischen Inter-coder-Reliabilität

Die Methodik setzt ein Sprachmodell als einzigen Produktionscodierer ein. Diese Designentscheidung erfordert eine transparente Positionierung gegenüber der klassischen Inter-coder-Reliabilität (ICR), wie sie in der quantitativen Inhaltsanalyse etabliert ist (Krippendorff, 2011; Cohen, 1960).

In der traditionellen Inhaltsanalyse zeigt die ICR, dass unabhängig geschulte menschliche Codierer dasselbe Codebuch auf dasselbe Material mit akzeptabler Übereinstimmung anwenden, typischerweise berichtet als Cohens κ oder Krippendorffs α . Das System stützt sich stattdessen auf zwei Masse: (a) die Wiederholungsmessung bei Temperatur 0 (Abschnitt 8.3), die beziffert, wie weit zwei Durchläufe derselben Sendung auseinanderliegen; und (b) den Vergleich mit den Entscheiden der Aufsichtsbehörde (Abschnitte 8.12 und 9), der zeigt, wie weit die Messung die Lesart einer unabhängigen Fachbehörde abbildet, die in einem förmlichen Verfahren entscheidet. Ein Vergleich mehrerer KI-Modelle liegt nur für die Fassung 3 vor (Abschnitt 8.5).

Die Wiederholungsmessung entspricht der Test-Retest-Reliabilität. Der Vergleich mit den Entscheiden der Aufsicht ist keine Inter-coder-Reliabilität: Er misst nicht, ob zwei Codierende zur selben Zählung kommen, sondern ob die Kennzahlen dort höher liegen, wo eine unabhängige Fachbehörde eine Verletzung festgestellt hat — an Entscheiden, an denen die Regeln entwickelt wurden. Auf der Ebene von Gesamtergebnissen liegt ein Vergleich mit einer manuellen Inhaltsanalyse vor: Für den deutschen öffentlich-rechtlichen Rundfunk kamen neun geschulte Codierende (Maurer, Kruschinski & Jost 2024) und das Instrument in einer früheren Fassung für alle untersuchten Formate zur selben Richtung der politischen Positionierung (Schlaepfer 2026e). Ein Abgleich auf der Ebene einzelner Beiträge ist das nicht.

Gleichwohl steht eine vollständige Mensch-gegen-Maschine-Reliabilitätsstudie — bei der geschulte menschliche Codierer das Codebuch unabhängig auf eine stratifizierte Stichprobe von Transkripten anwenden, mit Übereinstimmung berichtet als Krippendorffs α — noch aus; sie ist im Oktober 2026 angesprochen worden. Bis solche Evidenz vorliegt, stützen sich die Reliabilitätsansprüche des Systems auf die Wiederholungsmessung und den Vergleich mit den Entscheiden der Aufsichtsbehörde und nicht auf klassische ICR im strengen methodischen Sinne.

8.10 Externe Verankerung

Das Rahmenwerk ist in der Praxis benannter Aufsichtsbehörden verankert und nicht in abstrakten Qualitätsstandards. Seine Kriterien wurden anhand der Entscheidungen der Unabhängigen Beschwerdeinstanz für Radio und Fernsehen (UBI) entwickelt und kalibriert, deren Entscheide vollständig

veröffentlicht werden und beim Bundesgericht angefochten werden können; das BAKOM beaufsichtigt die übrigen Pflichten, und eine Ombudsstelle ist der Beschwerde vorgeschaltet.

Für die internationalen Kapitel sind die entsprechenden Behörden: Ofcom (Vereinigtes Königreich), KommAustria (Österreich), die Medienanstalten (Deutschland), ARCOM (Frankreich), AGCOM und GECA (Italien), CNMC (Spanien), Granskningsnämnden (Schweden), Medietilsynet (Norwegen), JSN (Finnland), Commissariaat voor de Media (Niederlande), RRTV (Tschechien), KRRiT (Polen), Médiateur (Ungarn), Coimisiún na Meán (Irland), RTÜK (Türkei), CRTC (Kanada), FCC (Vereinigte Staaten), ENACOM (Argentinien), ICASA und die BCCSA (Südafrika), BPO (Japan), die Korea Communications Standards Commission (Südkorea) und Roskomnadzor (Russland). Das Register hält für jede von ihnen fest, was sie prüfen, auf welchen Anlass hin, mit welcher Methode und in welcher Häufigkeit.

Der zentrale Befund dieser Erhebung ist für dieses Papier relevant: Ausserhalb Deutschlands, wo das Bundesverwaltungsgericht einen individuellen Anspruch bestätigt hat (6 C 5.24), gewährt keine der untersuchten Rechtsordnungen einen durchsetzbaren individuellen Anspruch auf ein ausgewogenes Programm. Die Aufsicht ist überwiegend beschwerdegetrieben, retrospektiv und einzelfallbezogen. Eine kontinuierliche, vergleichbare Messung konkurriert daher nicht mit der Aufsicht; sie liefert die empirische Schicht, die die Aufsicht nicht vorhält — wie das Sekretariat der UBI im September 2026 schriftlich mitteilte, gibt es neben den veröffentlichten Entscheiden keine Prüflistefäden, Checklisten oder Praxisübersichten; die Entscheide sind ihre einzige Datenbank, und systematisch ausgewertet wurde ihre Spruchpraxis nach ihrem Wissen nie.

8.11 Zuordnung zu klassischen Validitätstypen

Zur methodischen Transparenz ordnet die folgende Zuordnung die Validierungsdimensionen den vier klassischen Validitätstypen zu, die in der Kommunikationsforschung anerkannt sind:

Konstruktvalidität: Adressiert durch Abschnitt 8.7 (27 thematische Matrizen, gegeneinander geprüft), Abschnitt 8.1 (Sportkalibrierung als Nullhypothese), die externe Prüfung der Quellenampel-Dimension D3 an der wissenschaftlichen Publikationstätigkeit (Abschnitt 5) und die Einstufung jedes Kriteriums als eindeutig zählbar oder auslegungsbedürftig, gemessen an der Abweichung zwischen zwei Durchläufen (Abschnitt 4).

Kriteriumsvalidität: Adressiert durch die Abschnitte 8.12 und 9 (Vergleich mit 159 Entscheiden der UBI), eine externe Verankerung an den Entscheiden einer vom Forschungsteam unabhängigen Fachbehörde; die Regeln wurden an diesen Entscheiden entwickelt, eine Prüfung an Entscheiden, die bei der Entwicklung nicht vorlagen, steht aus.

Interne Konsistenz (Reliabilität): Adressiert durch Abschnitt 8.3 (Wiederholungsmessung: 0.71 Punkte im Gesamtscore und 1.29 in der Schwere unter der Fassung 4) und Abschnitt 8.4 (deterministische Berechnung aus veröffentlichten Formeln und zeitgestempelten Zählungen).

Externe Validität (Generalisierbarkeit): Adressiert durch Abschnitt 8.8 (ein Prompt, alle Länder), Abschnitt 8.2 (diskriminante Validität über Länder mit bekannten Unterschieden in der Pressefreiheit) und laufende internationale Erweiterung.

Klassische Intercoder-Reliabilität im strengen Sinne — Mensch-gegen-Mensch- oder Mensch-gegen-Maschine-Übereinstimmungskoeffizienten, berichtet als κ oder α — liegt noch nicht vor (siehe Abschnitt 8.9). Dies stellt die wichtigste verbleibende Lücke im Validierungsrahmen dar; eine solche Studie ist im Oktober 2026 angestossen worden, das Ergebnis steht noch aus.

8.12 Kalibrierung anhand von Aufsichtsentscheidungen

Das Rahmenwerk wurde anhand von 159 Sendungen aus 41 Formaten kalibriert, zu denen die Unabhängige Beschwerdeinstanz für Radio und Fernsehen (UBI) zwischen 1998 und 2026 eine materielle Entscheidung getroffen hatte — 35 gutgeheissene Beschwerden und 124 Abweisungen. Das Label ist nicht das Ergebnis des Dokuments, sondern die Schlussziffer für die tatsächlich gemessene Ausstrahlung: Eine Entscheidung kann eine Beschwerde über eine Morgensendung gutheissen und diejenige über die Hauptausgabe abweisen, und nur die im Korpus enthaltene Ausstrahlung zählt.

Weil das Format den Messwert stark mitbestimmt, wird innerhalb desselben Formats verglichen. In den 14 Formaten, in denen gutgeheissene und abgewiesene Sendungen vorliegen (34 gutgeheissene, 81 abgewiesene), liegen die gutgeheissenen im Mittel um 0.31 Punkte im Gesamtscore und um 0.66 Punkte in der Schwere über den abgewiesenen (Mittel der Differenzen je Format, gewichtet mit der kleineren der beiden Gruppen); in 9 der 14 Formate liegt das Mittel der gutgeheissenen höher, im Gesamtscore wie in der Schwere. Vertauscht man innerhalb jedes Formats die Bezeichnungen «gutgeheissen» und «abgewiesen» zufällig, entsteht ein so grosser Unterschied im Gesamtscore in rund sechs von hundert Fällen, in der Schwere in rund drei von hundert. Zu 30 der 35 gutgeheissenen Sendungen wurde eine abgewiesene als Gegenstück gemessen, in 29 Fällen aus demselben Format; im Gesamtscore lag die gutgeheissene zwölfmal über ihrem Gegenstück und 18-mal darunter, in der Schwere 17-mal darüber und 13-mal darunter. Für die einzelne Sendung trennt das Verfahren damit nicht; die Übereinstimmung zeigt sich erst im Mittel, in der Schwere deutlicher als im Gesamtscore. Ohne Trennung nach Format beträgt der Unterschied 0.11 Punkte im Gesamtscore und 0.42 in der Schwere. Nach Sprachregion getrennt liegen die gutgeheissenen in den acht deutschsprachigen der 14 Formate (24 gutgeheissene, 59 abgewiesene) um 0.67 Punkte im Gesamtscore und um 1.07 in der Schwere höher als die abgewiesenen, in den fünf französischsprachigen (9 gutgeheissene, 19 abgewiesene) um 0.60 Punkte im Gesamtscore und um 0.32 in der Schwere tiefer (Abschnitt 12); aus italienischsprachigen Formaten liegt nur eine gutgeheissene Sendung vor. Nicht in allen Formaten trennt das Verfahren: Im Gesamtscore liegen gutgeheissene Sendungen in vier Formaten von RTS und bei DOK nicht höher als abgewiesene, in der Schwere in drei Formaten von RTS, bei DOK und bei HeuteMorgen. Bei DOK und bei Temps présent ist die einzige gutgeheissene Sendung jene, bei der das Bundesgericht die Gutheissung aufhob. Formatspiegel — Mittelwerte pro Sendungstyp unter den von der Behörde geprüften Sendungen — werden in jedem Schweizer Bericht ausgewiesen, sodass ein Wert im Vergleich zu ähnlichen Sendungen statt an einer absoluten Skala abgelesen werden kann.

Drei weitere Rechnungen zeigen, wie weit das Ergebnis an einzelnen Entscheiden hängt. Das Bundesgericht hob neun der 35 Gutheissungen auf und stellte fest, dass die beurteilte Ausstrahlung die Programmbestimmungen oder das Sachgerechtigkeitsgebot nicht verletzte. Zählt man diese neun Sendungen zu den abgewiesenen, verschwindet der Unterschied im selben Format weitgehend: 0.09 Punkte im Gesamtscore und 0.29 in der Schwere; so grosse Unterschiede entstehen auch bei zufälliger Vertauschung der Bezeichnungen in rund jedem dritten beziehungsweise jedem fünften Fall. Das Verfahren bildet damit die Lesart der UBI ab, nicht den Ausgang vor Bundesgericht. Lässt man in der Lesart der UBI die 15 Sendungen weg, deren Entscheide die Zählregeln als Beispiel nennen, beträgt der Unterschied 0.40 Punkte im Gesamtscore und 0.80 in der Schwere. Fünf der abgewiesenen Entscheide betreffen mehrere Sendungen eines Zeitraums; gemessen ist je eine Sendung daraus. Ohne sie ändert sich der Unterschied kaum (0.31 Punkte im Gesamtscore, 0.63 in der Schwere). Einen Vergleich mit Entscheiden über einen ganzen Zeitraum enthält dieses Papier nicht.

Zwei Grenzen dieser Kalibrierung werden ausdrücklich benannt. Das Format wird zusammen mit dem Inhalt gemessen: Nachrichtensendungen und Debattensendungen unterscheiden sich systematisch unabhängig vom redaktionellen Verhalten. Und die Behörde entscheidet auf der Grundlage der Akten, nicht auf der Grundlage des Transkripts — ob eine Aussage korrekt wiedergegeben wurde oder ob eine Seite überhaupt um eine Stellungnahme gebeten wurde, ist im Untertiteltext nicht sichtbar, und ein Teil der gutgeheissenen Beschwerden beruht genau darauf. Ein tiefer Wert ist daher kein Freispruch und ein hoher Wert kein Schuldspruch.

8.13 Konstruiertes Testmaterial

Die Kalibrierung anhand realer Entscheidungen zeigt, ob das Rahmenwerk das Richtige misst; sie zeigt nicht, ob die Skala ihr oberes Ende überhaupt erreicht. Dafür wurden vier Sendungen gezielt so verfasst, dass sie jedes Kriterium verletzen, in den vier Schweizer Formaten, mit dokumentierten Gegenproben, bei denen nichts ausschlagen sollte. Unter der Fassung 4.12 erzielten sie 8.8, 8.7, 8.1 und 7.3; der Gesamtscore kann 10 nicht erreichen, weil mehrere Kriterien an ihren Maxima einander ausschliessen — eine Seite, die nie zu Wort kommt, kann nicht unterbrochen werden. Das Material wird nicht veröffentlicht und macht keine Aussage über reale Sendungen, Sender oder Menschen.

9. Prüfung an den Entscheiden der Aufsicht

Die wichtigste externe Prüfung ist der Vergleich mit den Entscheiden der Aufsicht; keine der genannten Institutionen ist mit dem Autor verbunden oder steht ihm nahe:

- UBI (Unabhängige Beschwerdeinstanz für Radio und Fernsehen): Die Ergebnisse werden mit UBI-Entscheidungen zu bestimmten Sendungen verglichen; dies ist der in Abschnitt 8.12 beschriebene Satz von 159 Sendungen, der sowohl zur Entwicklung der Regeln als auch zu ihrer Überprüfung dient. UBI-Entscheidungen sind rechtsverbindlich und werden in einem unabhängigen Verfahren getroffen. Wie weit die Scores des Systems den UBI-Ergebnissen folgen, zeigt, wie weit die Messung die Lesart der Aufsicht abbildet — an Fällen, an denen die Regeln entwickelt wurden.
- Bundesgericht: Die Urteile des Bundesgerichts dienen nicht als Massstab: Das Gericht prüft UBI-Entscheide als Rechtsmittelinstanz, nicht als Aufsichtsbehörde für Radio und Fernsehen (BGer 2C_406/2017 E. 1.2). In der Vergleichsmenge hob das Bundesgericht neun der 35 Gutheissungen auf; wie sich das auf den Vergleich auswirkt, zeigt Abschnitt 8.12.

Die UBI steht nicht unter der Kontrolle oder dem Einfluss des Forschungsteams; ihre Entscheide sind veröffentlicht, und die Labels der Vergleichsmenge sind aus ihnen abgelesen. Eine Prüfung an Entscheiden, die bei der Entwicklung der Regeln nicht vorlagen, steht aus; drei abgewiesene Fälle ausserhalb der Vergleichsmenge (b.870, b.888, b.917), im Oktober 2026 je einmal gemessen, ersetzen sie als Einzelmessungen nicht.

10. Internationale Abdeckung

Das archivierte Textkorpus umfasst 150 Länder in 70 Sprachen und besteht aus mehr als 26 Millionen Beiträgen (Untertiteldateien und Artikel). Neue Beiträge kommen jede Nacht hinzu; das analytische Rahmenwerk, die Kriterien und die Bewertungsregeln sind für jedes Land identisch, und dieselbe Prompt-Vorlage dient für alle. Die nachstehend aufgeführten 22 Länder sind diejenigen, für die das anwendbare Rundfunkrecht vollständig dokumentiert wurde, sodass Ergebnisse einer bestimmten nationalen Norm zugeordnet werden können:

- Schweiz (CH): SRF/RTS/RSI — Art. 4 RTVG
- Deutschland (DE): ARD/ZDF/DLF — §26 MStV
- Österreich (AT): ORF — §4 ORF-G
- Norwegen (NO): NRK — Medieansvarslov §4
- Niederlande (NL): NPO — Mediawet Art. 2.1
- Frankreich (FR): France Télévisions — Loi 86-1067 Art. 43-11
- Italien (IT): RAI — Contratto di Servizio Art. 6
- Kanada (CAN): CBC/Radio-Canada — Broadcasting Act s.3(1)(d)
- USA: PBS/NPR — 47 U.S.C. §396
- Irland (IRL): RTÉ — Broadcasting Act 2009 s.39(1)

- Vereinigtes Königreich (UK): BBC — BBC Charter Art. 6
- Spanien (ES): RTVE — Ley 17/2006 Art. 3
- Polen (PL): TVP — Art. 21 ust. 1 Ustawy o radiofonii i telewizji
- Finnland (FI): YLE — Laki Yleisradio Oy:stä 7 §
- Südafrika (ZA): SABC — Broadcasting Act s.6
- Tschechien (CZ): ČT — §2 Zákona o ČT / §31 Zákona č. 231/2001
- Argentinien (AR): TV Pública — Ley 26.522 Art. 3
- Südkorea (KR): KBS/MBC — 방송법 제6조
- Japan (JP): NHK — 放送法第4条
- Schweden (SE): SVT — Radio- och TV-lagen 5 kap.
- Türkei (TR): TRT — RTÜK-Gesetz Art. 8
- Ungarn (HU): MTVA — Mttv. §83

Drei weitere Länder verfügen über einen vergleichbaren Bestand an Analysen, sind aber nicht Teil der obigen Liste, weil ihre rechtliche Verankerung noch nicht ausformuliert wurde: Island (RÚV, 120 Berichte), Dänemark (DR, 108) und Russland (Perwy Kanal, 79). Elf weitere Länder existieren im Pilotstadium mit jeweils ein bis drei Berichten; ein einzelner Bericht ist ein Versuch, keine Beobachtung, und sie werden daher nicht als Abdeckung gezählt.

Jede Länderkonfiguration umfasst die vollständige politische Landschaft (Regierung, Opposition, Parteipositionen, CHES-Scores), den anwendbaren Rechtsrahmen und sprachspezifische Analyseanweisungen. Neue Länder können durch Definition dieser Parameter hinzugefügt werden, ohne den analytischen Kern zu verändern.

11. Begleitmethoden

Über die Kernanalyse nach 18 Kriterien hinaus haben wir mehrere Begleitmethoden entwickelt, die den analytischen Rahmen erweitern:

11.1 Rauscherkennung (Methode 14)

Ein Rahmenwerk zur Messung, ob Sender «mit Rauschen fluten» — d. h., ob wichtige Themen durch weniger relevante Berichterstattung verdrängt werden. Es verwendet 11 Dimensionen der Nachrichtenwichtigkeit (gewichtet nach demokratischer Relevanz, betroffener Bevölkerung, staatlicher Notwendigkeit und Sicherheit), um einen Rausch-Index zu erzeugen. Es stützt sich auf Forschung von Galtung/Ruge, Eilders, Colistra, Zaller, Franck und RAND. Detailliert beschrieben in Schlaepfer (2026b).

11.2 Emotionstrend-Analyse

Verfolgt emotionale Rahmungsmuster (Kriterien K1–K7) über Jahrzehnte der Sendegeschichte. Die erste Analyse umfasst mehr als 50 Jahre des schweizerischen öffentlich-rechtlichen Rundfunks und ermöglicht eine longitudinale Bewertung der redaktionellen Entwicklung.

12. Grenzen und Geltungsbereich

Ein niedriger Score ist kein Freispruch. Seit Fassung 4.0 wird ein Kriterium, dessen Auslösemenge null beträgt, als «nicht anwendbar» statt als null ausgewiesen, und der Mittelwert wird nur über die anwendbaren Kriterien gebildet. Eine Sendung kann daher einen niedrigen Score erhalten, weil nichts, was dieses Rahmenwerk misst, aufgetreten ist — reine Spielkommentare sind der deutlichste Fall: In drei von zehn gemessenen Spielen war nur eines von sechzehn Kriterien anwendbar (Abschnitt 8.1). «Nichts gefunden» und «nichts zu finden» sind unterschiedliche Aussagen, und der Bericht unterscheidet sie.

Der Gesamtscore ist die schwächste der ausgewiesenen Kennzahlen. Die Mittelung über sechzehn bewertete Kriterien begräbt den einzelnen schwerwiegenden Befund; Spitzenwert und Schwere existieren aus diesem Grund und sollten zuerst gelesen werden. Wie weit zwei Messungen derselben Sendung auseinanderliegen und wie genau Vergleiche zwischen Gruppen von Sendungen sind, ist in Abschnitt 8.3 beziffert. Das Rahmenwerk eignet sich daher für Aussagen über einen Sender über viele Sendungen hinweg, nicht für ein Urteil über eine einzelne Sendung.

Die folgenden Einschränkungen gelten für die aktuelle Methodik:

- **Reine Transkriptanalyse:** Das System verarbeitet Texttranskripte, nicht Audio oder Video. Visuelle Rahmung (Kamerawinkel, Bildauswahl, Grafiken (Texte auf T-Shirts und Mützen usw.) wird nicht erfasst. Dies ist eine bewusste Geltungsbereichsbeschränkung, die die Reproduzierbarkeit auf Kosten der Vollständigkeit erhöht.
- **Transkription:** Soweit Untertitel nicht von den Sendern stammen, entstehen sie durch maschinelle Transkription. Sie enthält Hörfehler, vor allem bei Namen und Mundart, und kann einzelne Zählungen verfälschen.
- **Einzelmodell-Produktion:** Die Produktion der Fassung 4 verwendet ein einzelnes Modell; ein Vergleich mehrerer Modelle liegt nur für die Fassung 3 vor (Abschnitt 8.5). Die Abweichung zweier Durchläufe (Abschnitt 8.3) ist die verbleibende Messunsicherheit jeder Einzelmessung.
- **Wissensstand-Grenze:** Das Weltwissen des Sprachmodells hat einen Trainings-Stichtag. Bei sehr aktuellen Ereignissen kann dem Modell Kontext fehlen, über den ein menschlicher Codierer verfügen würde. Das mit Zeitstempeln versehene Transkript kompensiert dies teilweise, indem es den sendungseigenen Kontext liefert.
- **Aktualisierung der politischen Landschaft:** Länderkonfigurationen (Parteipositionen, Regierungszusammensetzung) erfordern periodische manuelle Aktualisierungen. Zwischen den Aktualisierungen können Analysen auf leicht veraltete politische Konstellationen Bezug nehmen.
- **Sprachabdeckung:** Jede Analyse wird in der Sprache der Sendung verfasst; das archivierte Korpus umfasst 70 Sprachen. Bei Sprachen mit komplexer Morphologie oder begrenzten Trainingsdaten kann die Analysequalität geringfügig niedriger sein als bei den grossen europäischen Sprachen.
- **Sprachregion:** In den französischsprachigen Formaten trennt das Verfahren gutgeheissene und abgewiesene Sendungen nicht (Abschnitt 8.12). Ob das an den Zählregeln liegt, die an mehrheitlich deutschsprachigen Entscheiden entwickelt wurden (keine der 15 Sendungen, deren Entscheide sie als Beispiel nennen, ist französischsprachig), oder an der Gewichtung der UBI, ist offen; die Vergleichsmenge enthält nur neun gutgeheissene französischsprachige Sendungen, aus italienischsprachigen Formaten nur eine.
- **Vergleich mit menschlichen Codierern:** Der Vergleich mit rechtlich bindenden Entscheiden der Aufsichtsbehörde (UBI) verankert die Kriteriumsvalidität. Allerdings wurde eine klassische Interkodierer-Reliabilitätsstudie mit geschulten menschlichen Codierern, die das Codebuch unabhängig anwenden — ausgewiesen als Krippendorffs α —, noch nicht durchgeführt. Dies stellt den bedeutsamsten ausstehenden Validierungsschritt dar und ist als eigenständige Studie im Oktober 2026 angestossen worden; das Ergebnis steht noch aus. Die aktuelle Evidenz zur Reliabilität stützt sich auf die Wiederholungsmessung (Abschnitt 8.3) und den Vergleich mit den Entscheiden der Aufsichtsbehörde (Abschnitt 8.12).

13. Zugang für die Forschung

Die Methodik ist vollständig in diesem Papier und in den in Abschnitt 14 aufgeführten Begleitpapieren dokumentiert. Anfragen von Forschenden, Regulierungsbehörden und Institutionen werden individuell beantwortet:

- **Verifizierung:** Wenn eine spezifische Frage auftritt, können einzelne Analysen neu berechnet und die ihnen zugrunde liegenden Zählungen anhand des Transkripts überprüft werden.

- **Begleitpapiere:** Detaillierte Beschreibungen einzelner methodischer Komponenten werden als Arbeitspapiere veröffentlicht (siehe Abschnitt 14).
- **Musterberichte:** Illustrative Vollanalysen werden auf Anfrage zur Verfügung gestellt.

14. Literaturverzeichnis und Begleitpapiere

Begleitpapiere:

- Schlaepfer, D. (2026a). Systematic AI-Assisted Analysis of Public Broadcaster Impartiality: A Scalable Methodological Framework for Measuring Structural Bias in Public Service Media. SSRN Electronic Journal, Nr. 6688478.
- Schlaepfer, D. (2026b). Measuring Editorial Noise: A Retrospective Suppression Index for Public Broadcasting Content Analysis. SSRN Electronic Journal, Nr. 6733280.
- Schlaepfer, D. (2026c). The Source Traffic Light: A Six-Dimensional Credibility Framework for Systematic Source Assessment in Public Service Media. SSRN Electronic Journal, Nr. 6733880.
- Schlaepfer, D. (2026d). External Validation of a Multi-Dimensional Source Credibility Framework: Correlating Expert Ratings with Scientific Publication Output. SSRN Electronic Journal, Nr. 7326258.
- Schlaepfer, D. (2026e). Temporal Stability of Political Positioning in German Public Broadcasting: Convergent Evidence from Manual and Automated Content Analysis. SSRN Electronic Journal, Nr. 7344701.

Rechtliche Grundlagen:

- Bundesgesetz über Radio und Fernsehen (RTVG), Art. 4, 17, 24, 86, 89. SR 784.40.
- Länderspezifische Rundfunkgesetze gemäss Abschnitt 10.

Methodische Referenzen:

- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Colistra, R. (2012). Shaping and cutting the media agenda: Television reporters' perceptions of agenda- and frame-building and agenda-cutting influences. *Journalism & Communication Monographs*, 14(2), 85–146. <https://doi.org/10.1177/1522637912444106>
- Eilders, C. (2006). News factors and news decisions: Theoretical and methodological advances in Germany. *Communications*, 31(1), 5–24. <https://doi.org/10.1515/COMMUN.2006.002>
- Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4), 51–58. <https://doi.org/10.1111/j.1460-2466.1993.tb01304.x>
- Galtung, J., & Ruge, M. H. (1965). The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of Peace Research*, 2(1), 64–90. <https://doi.org/10.1177/002234336500200104>
- Jolly, S., Bakker, R., Hooghe, L., Marks, G., Polk, J., Rovny, J., Steenbergen, M., & Vachudova, M. A. (2022). Chapel Hill Expert Survey trend file, 1999–2019. *Electoral Studies*, 75, 102420. <https://doi.org/10.1016/j.electstud.2021.102420>
- Krippendorff, K. (2011). *Computing Krippendorff's alpha-reliability*. University of Pennsylvania ScholarlyCommons.
- Maurer, M., Kruschinski, S., & Jost, P. (2024). *Fehlt da was? Perspektivenvielfalt in der deutschen Medienberichterstattung*. Johannes Gutenberg-Universität Mainz.

- Zaller, J. (1992). *The nature and origins of mass opinion*. Cambridge University Press.

Dank

Der Autor dankt Prof. Yasemin Giritli Inceoglu (Bilgi-Universität, Istanbul) für eine eingehende kritische Durchsicht dieses Manuskripts. Ihre Beobachtungen zum Verhältnis von rechnerischer Konsistenz und klassischer Inter-coder-Reliabilität, zum epistemologischen Status der analytischen Kategorien, zu den Anforderungen an die Operationalisierung der Quellenampel und zur Zuordnung des Validierungsrahmens zu etablierten Validitätstypen der Kommunikationsforschung haben die Überarbeitung zur Fassung 3.4.1 wesentlich geprägt. Allfällige verbleibende Mängel liegen allein in der Verantwortung des Autors.

15. Versionstabelle

Fassung	Datum	Was sich geändert hat
3.4.1	Juli 2026	Erste vollständige Methodenbeschreibung auf SSRN: fünf Analyseschritte, 15 Kriterien, Quellenampel, Dreifachvalidierung.
3.5	September 2026	Einreichungsfassung für Blind Review (anonymisiert): redaktionelle Schärfung, interkulturelle Operationalisierung der Quellenampel, Zuordnung zu klassischen Validitätstypen. Keine Änderung an Kriterien oder Bewertung.
3.6	26. September 2026	17 Kriterien (K9 aufgehoben, K16 Behutsamkeit und Menschenwürde nach schwedischem Vorbild, K17 Vorverurteilung, K18 Quellenbelegung hinzugefügt); jedes Kriterium wird anhand einer expliziten Zählregel bewertet, lokal berechnet; Spitzenwert und Schwere neben dem Gesamtscore; zwei Messarten für K1, K5, K10 und K15, wobei Unser Mass zuerst angegeben wird; Kalibrierung anhand von 165 Aufsichtsentscheiden; Sportkalibrierung mit Messwerten; Aufsichtsbehörden namentlich genannt; Korpuszahlen aktualisiert (150 Länder, 69 Sprachen, mehr als 23 Millionen Einheiten); Konverter 3.9, Freigabe-Gate und Softfact-Sweep.
3.7	18. Oktober 2026	Abschnitt 9 neu gefasst: Validierung anhand der Entscheide der Aufsichtsbehörde und der BAKOM-Daten; die Rolle des Bundesgerichts abgegrenzt (Rechtsmittelinstanz, nicht Aufsichtsbehörde; Urteil 2C_406/2017 E. 1.2). Zusammenfassung, Abschnitte 8.9 und 8.11 und Grenzen der Studie entsprechend angepasst.
3.8	24. Oktober 2026	Stand Prompt 4.14: keine Blöcke Hardfacts und Softfacts mehr, Gesamtscore als Mittel aller anwendbaren Kriterien, jedes Kriterium als eindeutig zählbar oder auslegungsbedürftig eingestuft; K5 heisst Zahlendarstellung. Abschnitte 8.12 und 9 auf den Vergleich mit 159 Entscheiden der UBI ausgerichtet, mit Kennzahlen innerhalb desselben Formats (+0.31 im Gesamtscore, +0.66 in der Schwere), mit Gegenrechnungen (Lesart des Bundesgerichts, ohne die Beispielfälle der Zählregeln, ohne die Entscheide über einen Zeitraum) und getrennt nach Sprachregion; Grenzen der Studie um Transkription und Sprachregion ergänzt; Abschnitt 8.5 als Modellvergleich

		der Fassung 3; Wiederholungsmessung nach der Rechenart 4.14 (0.71 und 1.29); Abschnitt 8.1 mit zehn ganzen Spielübertragungen der SRG unter 4.14 neu gemessen. Korpus 70 Sprachen, mehr als 26 Millionen Beiträge. Abschnitt 7.3 beschreibt, wie die Auswertung den Score aus den Zählungen rechnet; Abschnitt 11.2 (Abstimmungsprognose) entfällt. Kurzfassung in einfacher Sprache, Lizenzvermerk, externe Prüfung der Quellenampel-Dimension D3, Quellenampel-Kategorien sowie K6, K10 und K14 nach dem Prompt 4.14; Abgleich mit menschlichen Codierenden angestossen (Oktober 2026). 18 Kriterien: 17 Qualitätskriterien und als 18. (K19) die politische Einordnung der Parteien; die Quellenampel ist ein Werkzeug der Expertenauswahl, kein eigenes Kriterium. Danksagung.
--	--	---